Harald H. Zimmermann, Saarbrücken

Mehrsprachiger Zugang zu textorientierten Datenbanken
(GAL-Jahrestagung 27.-29.9.1990, Bonn)

KURZFASSUNG
Der monolinguale Zugang zu Textdatenbanken ist nach wie vor geprägt von eher primitiven maschinellen Verfahren; die Texterschließung erfolgt entweder traditionell durch intellektuelle Indexierung (Deskriptorvergabe) oder durch Klassifikation. Da die meisten Datenbanken englischsprachig sind, auch der Markt eher im englischsprachigen Raum zu finden ist, besteht kein großer Bedarf an fremdsprachigen oder mehrsprachigen Erschließungs- oder Zugangssystemen. Insgesamt gesehen ist der Nutzer aber durch ständige Zeichenmanipulationen beim Suchvorgang stark belastet.

Es wird anhand von Forschungs- und Entwicklungsprojekten gezeigt, welche praktischen Möglichkeiten und Funktionen sich heute anbieten. Da diese Funktionen den Benutzer heute schon extrem von trivialen Techniken entlasten, wird für die Zukunft eine deutliche Trendwende zugunsten einer stärkeren Integration sprachpraktischer Verfahrensweisen prognostiziert.

GLIEDERUNG

1. Rahmenbedingungen bei textueller Datenbankrecherche

Unter einer textorientierten Datenbank werden im folgenden Informationssysteme verstanden, bei denen ein wesentlicher Teil der gespeicherten Daten aus frei formulierten Texten besteht.

(1) Es handelt sich dabei einerseits um Referenzdatenbanken bzw. Literaturdatenbanken: Hier gilt die Suche einem Literaturnachweis. Das Ergebnis der Suche wird in Form eines Belegstellennachweises geliefert, wobei neben dem Autor, dem Hinweis auf Erscheinungsort, Jahr usf. auch der Titel und ggf. eine Kurzfassung sowie in der Regel Stichwörter, Schlagwörter oder Deskriptoren angegeben werden.

(2) Ein zweiter Bereich sind die sog. Volltextdatenbanken. Hier steht die textuelle Information im Mittelpunkt. Beispiele sind Patentdatenbanken, Rechtsdatenbanken und Nachrichtendatenbanken. Nicht immer ist der gesamte Text des Originals gespeichert; bei Patenten oder Urteilen sind beispielsweise häufig nur die zentralen Materialien (Patentansprüche bzw. Kurzfassung; richterliche Leitsätze) verfügbar.

Daneben ist die technische Rahmensetzung von Bedeutung. Einerseits ist dabei zwischen den Zugangsmöglichkeiten zu unterscheiden:

(3) Bei den Online-Datenbanken erfolgt der Zugang über das öffentliche Netz (klassisches Telefonnetz, DATEX-P, ISDN). Dies bedeutet bislang (da ISDN, das relativ schnelle neue digitalisierte Datenvermittlungsnetz, noch kaum verfügbar ist) eine erhebliche Beeinträchtigung der Datenübertragungsgeschwindigkeit für den "Normalkunden": Mit 1.200 Bit/sec. benötigt die Übertragung einer DIN-A4-Seite auf dem Netz ca. 20 Sekunden. Die Systeme selbst bauen - EDV-technisch betrachtet - auf relativ alten Verfahren und Strategien auf (z.T. schimmert bei den Formaten die Lochkartentechnik noch durch) und müssen zudem ein weltweites Klientel bedienen, das z.T. selbst nur über einfache Zugangstechniken verfügt. Natürlich ist auch hier die Entwicklung nicht stehengeblieben.

(4) Bei den Inhouse-Systemen (dazu rechne ich auch die CD-ROM-Lösungen) sieht es schon etwas besser aus, da man die Daten entweder am PC vor Ort auf Festplatte oder eben auf der CD-ROM und ähnlichen Speichern verfügbar hat, die Datenübertragungsgeschwindigkeit in einem vernetzten lokalen System ist zudem erheblich größer. Einerseits bieten heute schon viele Hersteller von Online-Datenbanken parallel CD-ROM-Material an, andererseits kommen vor Ort mit der Bürokommunikation weitere textuelle Materialien wie Briefe, Berichte, Protokolle hinzu, die abgelegt und recherchiert werden müssen oder sollen.

Die bestehenden technischen Zugangsverfahren sind bei den textuellen Teilen v.a. in Referenzdatenbanken - mit wenigen Ausnahmen - losgelöst von den unterschiedlichen natürlichsprachigen Rahmenbedingungen entwickelt worden, wie überhaupt der Zugang über sprachliche Repräsentationsformen (d.h. den Wörtern aus dem Titel oder Abstract, bei Volltextdatenbanken ist das etwas anders) nur eine Nebenrolle spielt.

Die Suche mit einer Klassifikation, einer Graphik (Beispiel Chemie, Patente), mit einer "systematisierten" sprachlichen Begriffsrepräsentation (dem Deskriptor), mit dem Autor bzw. einem sonstigen formalen Merkmal wie Publikationsjahr, -ort usf. wird (da zudem technisch einfacher) weitaus stärker gestützt. Mehr darüber in den folgenden Abschnitten.

2. Problemstellungen bei der Texterschließung

Zunächst sollen die linguistischen Problemstellungen bei der Erschließung von Texten für textorientierte Datenbanken - ohne Anspruch auf Vollständigkeit - kurz angeführt werden.

Eine Texterschließung ist dabei nicht losgelöst vom späteren Suchzweck oder auch den Suchverfahren zu sehen, sondern hat eine darauf bezogene dienende Funktion. Sie stellt also immer zugleich eine Prämisse oder Erwartenserwartung dar, wie und in welcher Absicht ein späterer Informationssucher (der ja ein Problem lösen, eine Wissenslücke schließen möchte) sich der Systemhilfen bedient. Selbst wenn es beispielsweise gelänge, Texte (z.B. Aufsätze) auf ihre wesentlichen Inhalte zu verdichten (Abstracting), müßte man entsprechend unterschiedliche Aspektierungen vornehmen, um auf dem Weg der Verdichtung nicht relevante Informationen zu verlieren.

Die Texterschließungsverfahren im herkömmlichen Sinne - auf die ich mich wegen der praktischen Relevanz in bestehenden Textdatenbanken im folgenden beziehe - lassen sich grob wie folgt unterscheiden:

(1) Es gibt keine Texterschließung, sieht man von sog. Stoppwortverfahren ab, bei denen hochfrequente Wortformen von der späteren Suche ausgenommen werden. Hier müssen entweder der Benutzer oder das System beim Suchvorgang (Retrieval) entsprechend mehr tun (vgl. den Abschnitt 3).

Als "Ersatz" kann man natürlich die Fachgebietsklassifikation anführen, beispielsweise die Internationale Patenklassifikation.

(2) Es gibt eine intellektuelle Erschließung (z.B. in Form eines natürlichsprachigen Abstracts, das aber wiederum "Text" darstellt), wobei ein Indexierer meist anhand eines Thesaurus zur späteren Suche qualifizierte (Fach-)Wörter bereitstellt, die sog. Deskriptoren.

(3) Es gibt eine maschinelle Erschließung, die die menschliche Indexierung (Fachgebietserschließung, Deskribierung) "simuliert". Solche Systeme sind aber heute noch kaum in der Anwendung (Beispiel: Das Verfahren AIR von G. Lustig und die Anwendung AIR-PHYS am Informationszentrum Karlsruhe).

(4) Es gibt eine maschinelle Erschließung, die bei stark flektierenden und komponierenden Sprachen die Wortformen eines Textes auf Grundformen und (sinnhafte) Bestandteile zurückführt. In Deutschland ist dies (für die deutsche Sprache) für praktische Anwendungen in Verbindung mit GOLEM durch das von SIEMENS entwickelte System PASSAT gut gelöst.

Allgemein ist festzuhalten, daß international gesehen bei der Freitexterschließung über die Bereitstellung von im Text auftretenden Wortformen für die spätere Suche nicht allzu viel "erschlossen" wird. Dennoch hat man die Problematik sehr wohl erkannt. Daher werden für das spätere Retrieval zumindest noch einige Zusatzinformationen bereitgestellt, wie etwa die Position eines Wortes im Text, die Satz- oder Absatzgrenzen u.ä.m.

Ein Grund für diese Vorgehensweise ist sicherlich das Nicht-Vorhandensein praktisch nutzbarer linguistischer Erschließungsverfahren für alle in einer Textdatenbank vorkommenden Sprachen, wobei festzuhalten ist, daß der Datenbankmarkt nach wie vor von englischsprachigem Material dominiert wird, Englisch sich zudem als internationale Kommunikationssprache etabliert hat ("renommierte" deutsche Wissenschaftler publizieren fast nur noch in Englisch) und Englisch - äußerlich betrachtet - kaum Probleme beim Retrieval bietet.

3. Problemstellungen beim Textretrieval

Der Zusammenhang zwischen Texterschließung und -retrieval wurde schon festgestellt. Festzuhalten war zugleich, daß in aller Regel (GOLEM/PASSAT sind die Ausnahme) - sieht man von der Klassifikation und traditioneller Deskriptorvergabe ab - die "Suchlast" auf den Retrievalteil eines Datenbanksystems verlegt ist.

Bei der Suche mit Wörtern, die in einem Text (Titel, Abstract, Volltext) vorkommen können, werden dem Anwender - neben der Suche und Verknüpfung "ganzer" Wörter (= Wortformen) folgende Hilfen angeboten (auch hier nur das Wichtigste):

(1) Die Trunkierung: Wortformen können (meist am Wortende) für die Suche abgeschnitten werden, wenn sich danach alternative Zeichenketten (Endungen) ergeben. Hierfür wird

ein Sonderzeichen verwendet, wobei auch "Längenbegrenzungen" möglich sind. KIND$ könnte z.B. die Formen KIND, KINDES, KINDE, KINDER, KINDERGARTEN, KINDISCH usf. umfassen, bei KIND$2 würden nur KIND, KINDES, KINDER, KINDE "gesucht".

(2) Die Abstandsmarkierung: Hiermit wird - für durch Zwischenraum getrennte Wortformen - eine Möglichkeit geboten, komplexere Mehrwortgruppen von Zufallsbeziehungen in einem Text zu unterscheiden. Mit der Markierung JURISTISCH$ ADJ3 PERSON$ ließen sich beispielsweise Verbindungen wie JURISTISCHE PERSONEN, aber auch JURISTISCHE UND NATÜRLICHE PERSONEN noch "finden".

Bereits dies zeigt deutlich, welche Tricks dem Nutzer zugemutet werden, um sich einerseits vor unliebsamen Treffern zu schützen und andererseits mehr relevante Texte zu erhalten.

Wie wohltuend sich PASSAT/GOLEM von dieser Problematik unterscheiden, wird jedem deutlich, der mit weitergehenden Verfahren wie z.B. der sog. Linkstrunkierung arbeitet. Dies führt in der Regel zu langen Rechenzeiten; um diese zu vermeiden, sind einige Systeme dazu übergegangen, die Wortformen zusätzlich rückläufig zu speichern und bei der Linkstrunkierung die invertierten Formen rechts zu trunkieren.

Wenn man diese Praxis betrachtet, die sicherlich noch dieses Jahrzehnt mehr oder weniger unbeschadet überstehen wird, so kann man eigentlich nur staunen, wie wenig die Linguisten - v.a. die Computerlinguisten - bislang hier Hilfestellung gegeben haben. Es ist nicht zuletzt der Mangel an qualitativ hochwertigen Funktionen der Grundformenermittlung für möglichst alle gängigen Sprachen, der heute die vielen Tausend Nutzer von Textdatenbanken tagtäglich mit technischem Schnickschnack belastet.

Auf einen Problemkreis, der bei den Perspektiven eines "intelligenten Retrieval" eine Rolle spielt, die Auswertung einer natürlichsprachigen Suchanfrage, werde ich im Zusammenhang mit MITI/IMIS (Abschnitt 6) näher eingehen.

Zum engeren Bereich des Vortrags, dem mehrsprachigen Zugang zu Textdatenbanken, gibt es so gut wie keine praxisrelevanten Entwicklungen. Bereits Mitte der 80er Jahre war jedoch an der Universität Tsukuba mit Hilfe des von Nagao entwickelten Übersetzungssystems Englisch - > Japanisch ein Test unternommen worden, bei dem bei der Recherche in einer englischsprachigen Datenbank (es handelte sich m.W. um die INSPEC-Datenbank) mit einem Begriff in Japanisch gesucht wurde; bei einem Treffer wurde das Abstract unmittelbar "roh" ins Japanische übersetzt (unbekannte Wörter blieben in Englisch stehen).

Damit ist eigentlich der wesentliche Weg zukünftiger Entwicklungen schon vorgezeichnet:

(3) Suche in einer fremdsprachigen Datenbank mit Begriffen aus der Mutter(fach)sprache;

(4) Übersetzung der Ergebnisse aus einer Fremdsprache in die Sprache des Benutzers.

Der zuletzt angesprochene Teilaspekt ist ungleich "kritischer", da einerseits eben der "Lückeneffekt" auftreten kann, andererseits auch Falschübersetzungen (von Stilfragen gar nicht zu reden) auftreten können. Demgegenüber steht der Vorteil, dass die Kosten für die Übersetzung selbst unmittelbar dem aktuellen Benutzer angelastet werden können und dementsprechend auch am aktuellen Bedarf gemessen werden.

Besser - wenn auch teurer - ist das Verfahren, das beim Saarbrücker Übersetzungsmodell STS angewendet wurde: Hier wurden ganze Datenbanktitel - intellektuell postediert - übersetzt. Damit ist eine gute Basis für die Recherche (hier in Deutsch oder Englisch) gegeben. Der Test wurde u.a. an der Datenbank ICONDA des Informationszentrums Raum und Bau (IRB), Stuttgart, durchgeführt. Diese Datenbank wird inzwischen zweisprachig weitergeführt.

4. Linguistisch basierte Lösungen und Lösungsansätze

Ich werde im Folgenden die linguistischen Fragestellungen behandeln, wie sie sich aus meiner Sicht darstellen, die eher praxisorientiert ist. Dabei möchte ich gleich der Beurteilung vorbeugen, als sei der Ansatz theoretisch unbrauchbar/unbrauchbar oder wenig ergiebig/ergiebig. Es geht mir nicht darum, ein linguistisches Modell zu entwerfen oder anzuwenden. Dazu sind die Fragestellungen bereits viel zu primitiv und heterogen. Keine der später angesprochenen Problemlösungen ist in sich durchgängig und bildet hundertprozentig die Sprachäusserungen in ein geschlossenes Sprachsystem ab. Ziel ist die Entwicklung brauchbar-praktischer Verfahren, die dem Nutzer (hier in der angesprochenen Umgebung) weiterhelfen, ohne dass er sich ständig technischer Tricks bedienen muss.

Ich spreche daher bewusst im folgenden von (mehr oder weniger) linguistisch basierten Lösungen.

(1) Die Flexionsmorphologie: Schon jede Schulgrammatik beschreibt die Typologie der Flexionsmuster, jedes bessere Wörterbuch einer Sprache zeigt die Merkmale zu Genus, Numerus, Kasus usf. an grossen Datenmengen auf. Es gibt für die gängigen Sprachen wie Deutsch, Englisch, Französisch, Italienisch praktisch keine ernsthaften Probleme, Verfahren zu entwickeln, die aus einer Grundform sämtliche Flexionsformen generieren und umgekehrt von einer Textwortform zu der oder den möglichen Grundform(en) gelangen. Über Sonderfälle (wie z.B. die Frage der Steigerungsfähigkeit eines bestimmten Adjektivs, die mögliche Pluralbildung eines Substantivs kann man streiten, doch ist dies für die Applikation in den o.a. Bereichen unerheblich.

(2) Die Homographie und Polysemie: Die Differenzierung von Wortklassenhomographie und bedeutungsbezogener Mehrdeutigkeit ist v.a. für die Texterschließung keineswegs trivial. Ein erster Schritt ist jedoch zunächst einmal die Entwicklung von Verfahren zur Identifikation solcher Phänomene in einem Text. Dies bedeutet, dass elektronische Wörterbücher verfügbar werden müssen, die diese Differenzierungen beschreiben, auf denen dann Identifikationsverfahren aufbauen, um die potentiellen Lesarten anzuzeigen. Dies ist letztlich eine Fleißarbeit.

    Die Auflösung von Lesartenvarianten unter Berücksichtigung des rein sprachlichen Kontexts (wobei man schon streiten kann, wo die Sprachsemantik aufhört und das Weltwissen

anfängt) hat sich als wenig praktikabel erwiesen. Umgekehrt gibt es kaum elektronische sprachverarbeitungsbasierte Verfahren (und wenn es sie gibt, sind sie sehr langwierig und zudem für beliebige Bereiche kaum anwendbar), die solche Lösungen anbieten.

(3) Die Identifikation von Mehrwortbegriffen: Die Erkennung von begrifflich zusammengehörigen, an der Sprachoberfläche durch Zwischenraum getrennten Wörtern bedarf weitläufiger lexikalischer oder statistischer Verfahren. Dazu ist festzuhalten, dass gelegentlich auch hier noch Mehrdeutigkeiten (also Lesartenvarianten) auftreten, dass dies aber sprachpraktisch gesehen weniger problematisch - weil selten - ist. Man könnte der Ansicht sein - besonders wenn man das Deutsche mit der Diskontinuität von Verbzusätzen betrachtet -, dass dies nur über eine Satz- oder Textanalyse möglich ist.

Das Problem dabei ist nur, dass es zwar viele Ansätze zu einer Satzstrukturanalyse gibt, die Verfahren aber entweder wegen ihrer Komplexität und Restriktionen oder aber wegen eines unzureichenden Ausbaus oder aber wegen ihrer zu geringen Robustheit gegenüber den realen Texten (mit Satzbaufehlern) oder schließlich wegen der mangelnden Geschwindigkeit nicht verfügbar sind. Die einzige Ausnahme scheinen die maschinellen Übersetzungssysteme zu bilden, die im Grunde ja ähnliche Probleme haben. Aber hier fehlt es z.Z. noch an Schnittstellen zu den Retrievalsystemen.

Im Folgenden soll an einigen praktischen Themen gezeigt werden, wie ich selbst versucht habe, eine Brücke zwischen linguistischen Vorstellungen und praktischen Anwendungen zu bauen. Dabei kommt mir natürlich die Erfahrung in allen möglichen Bereichen (u.a. in den Bereichen der traditionellen Lexikographie, der elektronischen Syntaxanalyse, der maschinellen Übersetzung, der automatischen Indexierung) zugute, die in der grundlagen- und anwendungsorientierten Forschung in den letzten 25 Jahren gesammelt wurden.

In allen vorgestellten Fällen handelt es sich um Projekte und Aufträge, die zumindest bei meinen Teilaufgaben nicht an der Universität und auch nicht an meinem Forschungsinstitut abgewickelt werden, sondern über die SOFTEX GmbH. Daran kann man v.a. eines erkennen: Zielsetzung ist es, nicht nur Verfahren und Funktionen prinzipiell zu erforschen, sondern sie - soweit sie erfolgreich sind - am Markt einzusetzen. Dass dies mit z.T. hohen Investitionen und einem langen Atem verbunden ist, führt einerseits dazu, dass man etwas vorsichtig plant, andererseits auch zu Schwachstellen, die mit etwas mehr finanziellen Mitteln durchaus hätten vermieden werden können.

5. Das EUREKA-Projekt EURO-TRIANGLE

Bei diesem Projekt geht es eigentlich um einen sog. Übersetzungsarbeitsplatz (die Betonung liegt bei "ungs"). Es werden dabei einerseits Hilfen zur Erstellung fremdsprachiger Texte entwickelt, andererseits wird eine Schnittstelle zum "Teletranslating" aufgebaut. Partner sind die SOFTEX GmbH, die Wiener Firma CIB, das Ludwig-Boltzmann-Institut, Wien (Federführung) und die GACHOT S.A., Frankreich (Bereitstellung von SYSTRAN als Host für die Fernübersetzung).

Der Beitrag von SOFTEX ist einerseits der Aufbau eines weitgehend konsistenten dreisprachigen elektronischen Wörterbuchs (Deutsch / Französisch / Englisch; Allgemeinwortschatz und Technik-Basiswortschatz), andererseits die Bereitstellung von Schnittstellen zur Übersetzung lexikalischer Einheiten. LIT / CIB entwickeln auf der Grundlage weiterer Funktionen einen Editor, der

es erlaubt, während der Texterstellung einen fremdsprachigen Begriff zu recherchieren und in den Text zu insertieren.

Das Projekt ist auf vier Jahre geplant; bereits das erste Jahr - das gerade beendet ist - hat zu einem Vorläuferprodukt geführt (PRIMUS-TRI), das den Editor mit Rechtschreibkorrekturfunktionen und Schnittstellen zwischen dem Korrektur-Editor und verschiedenen Textsystemen umfasst. Für die nächste Phase sind die Einbringung von Übersetzungshilfen in eine Retrievalumgebung sowie die Integration spezifischer Posteditionshilfen zur maschinellen Übersetzung vorgesehen (mit paralleler Textführung).

Ein wesentliches Element der SOFTEX-Entwicklungen bei EURO-TRIANGLE ist zudem der Ausbau des bestehenden Lexikonsystems unter Realisierung geeigneter Datenstrukturen (auch für große Datenvolumen) und eine weitergehende Differenzierung der Schnittstellen. Das Projekt wird anteilig vom BMFT finanziert.

6. Das IMPACT-Projekt MITI-IMIS

Die Fragestellungen des zu anteilig von der EG mitfinanzierten Forschungsprojekts MITI / IMIS sind in einigen Punkten verwandt mit den Themen von EURO-TRIANGLE. Im Mittelpunkt von MITI / IMIS, an dem u.a. die englische Firma TOME Associates, die Universität Toulouse, die GMD (Darmstadt) und die Firma EUROBROKERS (Luxemburg) beteiligt sind, steht die Entwicklung eines sog. "intelligenten" (multilingualen) Interface zu Informationsbanken. Der Nutzer soll seine Suchanfrage weitgehend unabhängig von der Zugangssprache der Datenbank formulieren können, das System "übersetzt" diese Formulierung in die notwendige Form des jeweiligen Hostsystems.

Hierbei werden elektronische Wörterbücher zu den vier Sprachen (Deutsch, Englisch, Französisch und Spanisch) entwickelt, wobei der Schwerpunkt einerseits in der Integration des Materials des sog. ROOT-Thesaurus liegt (ein Technik-Wörterbuch der British Standard Organisation). Einen weiteren Schwerpunkt der lexikalischen Entwicklungen bildet der Bereich Umweltschutz.

Die Aufgabenstellung von SOFTEX bezieht sich auf die eher funktionale Unterstützung einer mehrsprachigen Suche in Datenbanken. Ziel ist die Entwicklung bzw. Integration folgender Teilfunktionen:

(1) Erkennung und ggf. Korrektur von Rechtschreibfehlern
(2) Ermittlung möglicher Grundformen
(3) Ermittlung von Übersetzungsäquivalenten (auf Grundformenebene)
(4) Erzeugung von Flexionsformen aus Grundformen der jeweiligen Sprachen
(5) Automatische Ersetzung der Flexionsformen durch Trunkierungsangaben

Hinzu kommen folgende - über monolinguale Äquivalenzwörterbücher hergestellte - Wortbeziehungen:

(6) Morphologische Teilwortbeziehungen (v.a. für Deutsch)
(7) Derivationsrelationen zwischen Verb, Adjektiv und Substantiv
(8) Synonymbeziehungen

Insgesamt ist dabei ein spezifisches Inventar von rd. 50.000 Termini zu verknüpfen.

7. Das EG-Entwicklungsprojekt EQUITEXT

Ein großes Problem der elektronischen Sprachdatenverarbeitung ist die Inventarisierung geeigneter lexikalischer Materialien für elektronische Wörterbücher und Terminologien, insbesondere im Übersetzungsbereich. Es ist hier nicht der Ort, über die Verwendbarkeit bestehender (gedruckter) Wörterbücher zu sprechen. Andererseits verfügt v.a. die Kommission über große Mengen von Textmaterial, das in qualifizierten parallelen Übersetzungen in elektronischer Form vorliegt.

Auf der Grundlage einer Studie am Institut für angewandte Informationsforschung (IAI) wird z.Z. von SOFTEX im Auftrag der EG-Kommission der Prototyp eines textorientierten lexikalischen Erschließungssystems entwickelt (EQUITEXT). Seine wesentlichen Funktionen lassen sich wie folgt beschreiben:

(1) Umsetzung von Texten in ein neutrales Eingabeformat. Im Mittelpunkt des Tests, der zu den Sprachen Deutsch, Englisch und Französisch durchgeführt wird, steht die Einbringung von jeweils rd. 1.000.000 laufenden Wörtern Text aus den Datensammlungen des Amtes für Veröffentlichungen (FORMEX-Formatierung).

(2) Segmentierung und Parallelisierung der kleinsten Textsegmente eines Dokuments. Segmente sind kleinste Teile eines Textes, die sich eindeutig in Quell- und Zielsprache entsprechen. Dies geschieht im wesentlichen mittels formaler Kennungen (Dokument, Absatz) und innerhalb der Absätze nach einem statistischen Verfahren auf der Basis von Satzlängen.

(3) Die weiteren Verfahrensschritte arbeiten sowohl statistisch als auch linguistisch.

(4) Die linguistische Komponente basiert i.W. auf Übersetzungswörterbüchern; v.a. zum Deutschen werden auch die Teilwortrelationen und Derivationen herangezogen. Hier ist es ein Teilziel, möglichst solche Wörter zu identifizieren, die bereits bekannt sind. Dabei werden auch lexikalisierte Mehrwortgruppen zu erkennen versucht. Für "bekannte" Wörter lässt sich zumindest eine Frequenz im Gebrauch verschiedener Lesarten ermitteln.

(5) Da die statistischen Verfahren von Auftretenswahrscheinlichkeiten abhängen, ist die Grundformenermittlung (Lemmatisierung) auch hier wichtiges Teilelement. Es wird dabei - vereinfacht dargestellt - nach folgendem Konzept vorgegangen: Als mögliche Übersetzungsäquivalente werden Wörter der Quell- und Zielsprache betrachtet, die fast immer im gleichen Segment auftreten. Ein wichtiges Teilproblem, das v.a. bei maschineller Übersetzung auftritt, kann hierbei zugleich mit behandelt werden: Die Ermittlung "fester" Mehrwortgruppen (bzw. die Übersetzung von Komposita des Deutschen).

Die funktionalen Komponenten zu EQUITEXT sind inzwischen (unter Nutzung der bei SOFTEX verfügbaren Lexika) so weit entwickelt, dass ab November der Großtest stattfinden kann.

Es bietet sich im übrigen an, die Ergebnisse unmittelbar in Textdatenbanken umzusetzen. Hierzu sind "nur" noch die Lücken zu schließen, die sich bei der Zuordnung von "Restwörtern" ergeben. Es entsteht auf diese Weise gleichsam als Seiteneffekt eine multilingual zugängliche Textsamm-

lung. Da bereits bei der Textumsetzung Fachgebiets- und Textsortenmerkmale mitgegeben werden (das spezifische Schema dafür ist noch in Entwicklung), lassen sich die Verfahren auch selektiert nach Fachgebieten und Dokumentarten anwenden.

8. Teletranslating mit SYSTRAN

EQUITEXT wurde von der EG-Kommission u.a. auch im Hinblick darauf entwickelt, die Arbeiten zur Erweiterung der elektronischen Wörterbücher des Übersetzungssystems SYSTRAN zu verbessern bzw. zu beschleunigen. Eine Teilaufgabe ist die Umsetzung der Ergebnisse in das Kodierschema von SYSTRAN (auf das in diesem Zusammenhang nicht weiter eingegangen werden kann).

Es ist einleuchtend, dass das beste Retrievalsystem - auch der Zugang über einen mehrsprachigen Thesaurus bzw. ein entsprechendes Lexikon nichts nützen, wenn der Datenbanktext anschließend in einer Sprache verfügbar ist, die der Nutzer nicht ausreichend beherrscht. Es gilt also, den Kreis zu schließen, indem in den Prozess des multilingualen Retrieval ein maschinelles Übersetzungssystem mit eingebunden wird.

Wesentliche Anforderungen dieser Datenfernübersetzung - neben der prinzipiellen Verfügbarkeit für das Sprachpaar - sind die Übersetzungsqualität, gekoppelt mit einer angemessenen Übersetzungsgeschwindigkeit. Als ein am Markt verfügbares Verfahren ist das Übersetzungssystem SYSTRAN in der Lage, den dabei herrschenden Anforderungen Rechnung zu tragen.

Im Idealfall verfügt der Datenbankhost bereits über einen direkten "Draht" zum Übersetzungsrechner. Dieses Konzept wird spätestens 1991 realisiert sein. Es geht aber heute schon in zwei Etappen: Nach dem Downloading eines Textes auf den lokalen PC werden die Daten an den Übersetzungsrechner geschickt. Hierbei lässt sich - um ein konkretes Beispiel zu nennen - über die SYSTRAN-spezifische Schnittstellensoftware EXPRESS eine Verbindung herstellen, der Text wird unter Angabe der gewünschten Zielsprache und ggf. des Fachgebiets über DATEX-P an den Übersetzungsrechner gesendet; die Rohübersetzung kommt auf Wunsch innerhalb weniger Minuten zurück auf den PC.

9. Ausblick

Was hier aus der Werkstatt eines Unternehmens der Sprachindustrie vorgestellt wurde, wird manchem Linguisten, der im stillen Kämmerlein oder auch im Großforschungsprojekt in der Grundlagenforschung arbeitet, als oberflächlich erscheinen. In der Tat kamen in der Darstellung viele Probleme etwas zu kurz, etwa die Frage der semantischen Differenzierung lexikalischer Einheiten, die Systematisierung der Komposition, die Satzstrukturanalyse.

Eines ist mir jedoch seit längerem klar: Es müssen endlich Werkzeuge geschaffen werden, die erste, deutlich erkennbare Schritte der Erleichterung in der Textverarbeitung und im hier vorgestellten spezifischen Fall des mehrsprachigen Zugangs zu Textdatenbanken bringen, und zwar so, dass sie vom Markt (das sind nicht nur die Nutzer, sondern auch die Hersteller solcher Systeme) akzeptiert und eingebunden werden.

Es hat sich außerdem gezeigt, dass dann, wenn einmal eine breite Materialbasis geschaffen ist, die nächsten Schritte leichter von der Hand gehen. Die SOFTEX verfügt inzwischen über eine reichhaltige Funktionssammlung. Hinzu kommen ausreichend große monolinguale Identifikationswörterbücher zu Deutsch, Englisch und Französisch (inkl. der Ermittlung potentieller Grundformen), zu Italienisch, Spanisch und Portugiesisch sind entsprechende Verfahren im Aufbau. Im Übersetzungsbereich ist es nicht sinnvoll - dies zeigt die bisherige Erfahrung von SOFTEX -, die Strukturen "gedruckter" Wörterbücher zu übernehmen: die EDV-Lösung verlangt eigene Strukturen und Differenzierungen.

Die nächste Stufe - die inzwischen schon eingeleitet wurde - besteht in der stärkeren und v.a. eindeutigen semantischen Differenzierung. Wer dahinter eine tiefergehende Einsicht vermutet, der sieht sich jedoch getäuscht: Es sind die bestehenden Systeme, etwa gerade SYSTRAN, die diese Differenzierung bereits kennen (und natürlich ausbauen): Fachgebietskennungen, semantische Merkmale sind praktisch und brauchbar, um die Qualität der Verfahren zu erhöhen. Semantische Kriterien oder auch einfach Differenzierungen werden ferner benötigt, um Äquivalenzbeziehungen (i.S. strenger Synonymie) für die automatische Generierung von Sprachpaarbeziehungen zu nutzen: Wenn es schon eine Äquivalenzbeziehung zwischen einem deutschen Quellspracheneintrag und einen französischen Zielspracheneintrag gibt und eine weitere zwischen dem gleichen deutschen Quellspracheneintrag und einem englischen Zielspracheneintrag besteht, so lässt sich daraus - trivialerweise - eine entsprechende Zuordnung der beiden Zielspracheneinträge ableiten. Ohne semantische Differenzierung (bzw. Fachgebietskennungen) ist dies aber bei mehreren Übersetzungsmöglichkeiten problematisch. Um die Arbeit bei der Entwicklung elektronischer Lexika zu ökonomisieren, ist eine semantische Differenzierung erforderlich.

Die elektronische Sprachverarbeitung steht dabei - dies sei abschließend erwähnt - in einem Spannungsfeld zwischen starker Formalisierung, die der direkten DV-Anwendung, d.h. den Algorithmen nutzt, und einer benutzerorientierten Anschaulichkeit, wo - etwa beim Blick ins elektronische Wörterbuch - unverständliche semantische Kategorisierungen eher stören. Wer möglichst viele Synergieeffekte beim Aufbau der Funktionen wie den Anwendungen erzielen will, muss auf dem schwierigen Weg der Systematisierung, der Applikationsanforderungen und der system- und benutzerseitigen Anforderungen wandeln lernen.

Die neunziger Jahre bringen dank der Verbreiterung der Ausbildung, der extrem gewachsenen Leistungsfähigkeit der Computertechnik, der Vor-Ort-Verfügbarkeit der Telekommunikationstechniken und der inzwischen aus den Kinderschuhen herauswachsenden linguistischen Verfahren eine Chance, leistungsfähige Funktionen zur Sprachdatenverarbeitung als selbstverständliche Instrumente in die Standardsoftware der Hersteller einzubringen.

(T11GLV1)